nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Improving Polygenic Prediction in Ancestrally Diverse Populations
Corresponding author name(s): Dr Hailiang Huang

Reviewer Comments & Decisions:

Decision Letter, initial version:

IMPORTANT: Please note the reference number: NG-A56583R-Z Huang. This number must be quoted whenever you communicate with us regarding this paper.

7th Sep 2021

Dear Dr. Huang,

Thank you for your message of 7th Sep 2021, asking us to reconsider our decision on your manuscript "Improving Polygenic Prediction in Ancestrally Diverse Populations". I have now discussed the points of your letter with my colleagues, and we would like to invite you to resubmit your revised manuscript for further consideration at Nature Genetics.

When preparing a revision, please ensure that it fully complies with our editorial requirements for format and style; details can be found in the Guide to Authors on our website (http://www.nature.com/ng/).

At this stage we will need you to upload:
1) a copy of the manuscript in MS Word .docx format.
2) The Editorial Policy Checklist:
https://www.nature.com/documents/nr-editorial-policy-checklist.pdf
3) The Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
(Here you can read about the role of the Reporting Summary in reproducible science:
https://www.nature.com/news/announcement-towards-greater-reproducibility-for-life-sciences-research-in-nature-1.22062)

Please use the link below to be taken directly to the site and view and revise your manuscript:

[REDACTED]

Please let me know if you have any questions.

Thank you very much.

All the best,

Catherine

Catherine Potenski, PhD
Chief Editor
Nature Genetics
1 NY Plaza, 47th Fl.
New York, NY 10004
catherine.potenski@us.nature.com
https://orcid.org/0000-0002-4843-7071

Author Rebuttal to Initial comments

NG-A56583

Improving Polygenic Prediction in Ancestrally Diverse Populations

Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Stanley Global Asia Initiatives, Lin He, Akira Sawa, Alicia R. Martin, Shengying Qin, Hailiang Huang, Tian Ge

We thank all reviewers for the insightful comments and constructive suggestions, and the editor for giving us the opportunity to revise the manuscript. We have carefully considered the points made by the reviewers and revised the manuscript accordingly. Specifically, we have made several major changes and additions to the manuscript:

(i) We have significantly expanded the simulation studies. In addition to the effect of varying genetic architectures (fraction of causal variants) and cross-population genetic correlations on the predictive performance of different polygenic prediction methods examined in the previous version of the manuscript, we have also assessed the impact of discovery GWAS sample size, SNP heritability, proportion of causal variants shared across populations, allele frequency and LD dependent genetic architecture, and the selection of hyper-parameters on the prediction accuracy.

(ii) We have expanded the quantitative trait analysis in the UK Biobank (UKBB) and Biobank Japan (BBJ) from 16 traits to 33 traits, and the analysis in the UKBB, BBJ and PAGE study from 7 traits to 14 traits. In addition, we have updated all BBJ GWAS to the latest version as described in Sakaue and Kanai et al. (*medRxiv*: https://doi.org/10.1101/2020.10.23.20213652). This newer version of BBJ summary statistics has an increased sample size and improved imputation quality for many traits.

(iii) We have added the Taiwan Biobank (TWB) as a new target dataset to assess the predictive performance of different PRS construction methods in an East Asian population. We have brought on a new co-author, Dr. Yen-Feng Lin, to the manuscript who contributed to the TWB analysis.

(iv) We have updated all LDpred results to LDpred2 (Prive et al. *Bioinformatics* 36, 5424-5431), reflecting the recent improvement in the LDpred algorithm.

(v) We have added LDpred2-mult and PRS-CS-mult to all simulations and real data analyses to provide a more comprehensive comparison across different polygenic prediction methods. In particular, a comparison between PRS-CS-mult and PRS-CSx demonstrates the benefits of jointly modeling multiple GWAS summary statistics across populations using a coupled continuous shrinkage prior.

We now provide a point-by-point response to specific reviewer concerns. Our response is preceded by a double-dash (--) and rendered in a blue typeface. Major changes to the text are highlighted in red in the updated manuscript. We believe that the manuscript has been substantially improved and hope it is now suitable for publication in *Nature Genetics*.

Reviewer #1:

This paper describes a new approach (PRS-CSx) to construct PRS under the cross-population setting. The method is an extension of PRS-CS (their previous work) which is a Bayesian framework with a continuous shrinkage (CS) prior on SNP effect sizes. And in PRS-CSx, the authors couple the genetic effects across populations via a shared continuous shrinkage prior, providing a flexible way to aggregate information from different populations. By applying to several simulations and real data, they have shown improved genetic risk prediction performance of PRS-CSx relative to alternative approaches. However, as to the analyses and results, I have a number of concerns.

1. It isn't clear to me if the improved performance of PRS-CSx is due to the usage of shared continuous shrinkage prior, or is simply due to the linear combination of more than one PRS constructed from more data. Since during the comparison with other multi-ethnic methods, the authors only compared PRS-CSx

with PT-meta and PT-multi which are both based on PT, it is worth checking the performance of LDPred-meta/LDPred-multi or PRS-CS-meta/PRS-CS-multi to be more persuasive about the advantage of having such a shared continuous shrinkage prior in the cross-population setting. And from another perspective, if I have two independent data from the same population, will use PRS-CSx be better or the same as using single PRS-CS directly on the combined data?

-- We thank the reviewer for this insightful comment and agree that a detailed investigation on the advantage of the coupled continuous shrinkage prior was missing. In this revised version of the manuscript, we have added LDpred2-mult and PRS-CS-mult to all simulations and real data analyses. In particular, a comparison between PRS-CS-mult and PRS-CSx quantifies the benefits of using a coupled shrinkage prior across populations. Our results showed that PRS-CSx consistently outperformed the prediction of PRS-CS-mult in non-European populations, although the amount of improvement varied across traits and was not always substantial. For example, when integrating UKBB and BBJ GWAS and predicting into the EAS population, PRS-CSx had a median relative improvement of 8.3% and 7.1% over PRS-CS-mult in the EAS and AFR populations, respectively (Figure 3). The improvement of PRS-CSx over LDpred-mult was greater, with a median relative increase of 10.5% and 22.2% in the EAS and AFR populations, respectively (Figure 3). We also note that, while this comparison helps to dissect the contributions from sample size increase vs. the use of the coupled shrinkage prior to prediction accuracy, PRS-CS-mult and LDpred-mult are not published methods *per se* and have not been used in any studies.

We chose to compare PRS-CSx against LDpred2-mult and PRS-CS-mult rather than LDpred2-meta and PRS-CS-meta because (i) The only modeling difference between PRS-CSx and PRS-CS-mult is the use of a coupled shrinkage prior. Comparing PRS-CSx with the “mult” methods thus provides a direct characterization of the benefits of the coupled prior. (ii) The LD pattern of a cross-ancestry meta-analyzed GWAS is a mixture of population-specific LD patterns, which is difficult to be appropriately modelled using current methods. Meanwhile, an accurate modeling of the LD patterns is critical to Bayesian polygenic prediction methods. In fact, we have observed convergence issues for LDpred2 when there is mismatch between the discovery GWAS summary statistics and the LD reference panel. We therefore do not recommend LDpred2-meta and PRS-CS-meta from a modeling perspective. The same recommendation was made by other recent studies in the literature (e.g., *medRxiv*: https://doi.org/10.1101/2021.01.19.21249483).

Regarding the last question, PRS-CSx is designed to flexibly model GWAS summary statistics from multiple populations where SNP effect sizes and/or LD patterns differ. For two or more GWAS conducted in independent samples from the same population, where effect sizes and LD patterns are expected to be highly concordant, a fixed-effect meta-analysis, whose modeling assumptions are fully satisfied in this scenario, is probably the optimal approach to combine the GWAS and maximize statistical power. We have clarified this in the revised manuscript. See Page 11, Line 400-404.

2. The authors provide nice figures (Figure 3) to show the improvement of PRS-CSx over other methods, but when it comes to individual traits (Table S4), the improvements are actually not statistically significant with overlapping confidence intervals for most traits (e.g. Red blood cell count, blood pressure, etc.).

-- Thanks for the comment. We agree that the improvement of PRS-CSx over alternative methods, in particular, LDpred2-mult and PRS-CS-mult, is trait-specific. While the prediction accuracy for many traits was substantially increased (e.g., when integrating UKBB and BBJ GWAS and predicting into the EAS population, PRS-CSx improved the prediction accuracy over PRS-CS-mult by >10% for 15 out of 33 traits), as the reviewer pointed out, the improvement can be small and non-significant for other traits. We think this primarily reflects the relatively low heritability and the lack of powerful non-EUR GWAS for many traits, rather than a weakness

of PRS-CSx. In fact, PRS-CSx provided improvement over alternative methods for the majority of the traits we examined (e.g., in the UKBB+BBJ EAS prediction, PRS-CSx was more accurate for 32 and 27 out of 33 traits than PRS-CS-mult and LDpred-mult, respectively), and the trend was highly significant (two-sided sign test P = 7.92×10^{-9} and 3.24×10^{-4}, respectively), supporting our conclusion that PRS-CSx improves the prediction accuracy over alternative methods in non-European populations. We have added some discussions around these points in the revised manuscript, and suggested that future research is needed to dissect the effects of potential factors (e.g., the comparative genetic architecture across populations and the sample characteristics of the target dataset) on the accuracy of cross-population prediction and to better understand the behavior of different prediction algorithms for individual traits. See Page 11, Line 406-411.

3. The model used to generate the simulation matches the authors' model exactly, where they assume the causal variants are the same across populations. They may also want to investigate the potential for bias when the model is misspecified, for example, when only a segment of the causal SNPs is shared between populations and when the heritability is different across populations.

-- Thanks for this helpful suggestion. As mentioned at the beginning of the response letter, we have significantly expanded the simulation studies in the revised manuscript. Specifically, to address the two questions here, we have added a new set of simulations where the proportion of shared causal variants across populations was reduced to 70% or 40% (Supplementary Figure 6), and another set of simulations where trait heritability was different across populations (h^2=0.5 and 0.25 in EUR and non-EUR populations respectively, and vice versa; Supplementary Figure 5). Our simulation results showed that the PRS-CSx model is robust to model misspecification and varying trait heritability across populations. See Page 6, Line 229-238; Supplementary Figures 5 and 6.

4. When applying the methods to quantitative traits for EUR target, even with around 100,000 EAS samples (BBJ) added to the UKBB, the improvement of prediction performance is still very limited. The authors explain that it is likely because the UKBB GWAS were already well-powered. To make this claim stronger, it is better to do more simulations to show the relations between sample size and prediction performance to prove that there exists such a threshold that makes the GWAS "well-powered".

-- In the revised manuscript, we have conducted a new set of simulations where we varied the sample size of the discovery GWAS with the ratio of the EUR vs. non-EUR GWAS sample sizes kept constant (50K EUR + 10K non-EUR; 100K EUR + 20K non-EUR; 200K EUR + 40K non-EUR; 300K EUR + 60K non-EUR), and a second set of simulations where we varied the ratio of the EUR vs. non-EUR GWAS sample sizes with the total sample size kept constant (120K EUR + 0K non-EUR; 100K EUR + 20K non-EUR; 80K EUR + 40K non-EUR; 60K EUR + 60K non-EUR). In the first set of simulations, integrating non-EUR GWAS with EUR GWAS using PRS-CSx provided limited improvement in prediction accuracy over PRS-CS trained on the EUR GWAS in the EUR target population, regardless of the GWAS sample size (Supplementary Figure 3). In the second set of simulations, adding 60K non-EUR GWAS to the 60K EUR GWAS clearly improved the prediction over single-discovery methods in the EUR population, and the improvement decreased as the ratio of the EUR vs. non-EUR GWAS sample sizes became larger (Supplementary Figure 4). These results suggested that when the target population is EUR, adding a small non-EUR GWAS provides limited gain in prediction power, but non-EUR GWAS can make an impact on the prediction when their sample sizes become more comparable to the EUR GWAS. In real data analysis, since the vast majority of non-EUR GWAS remained under-powered relative to the EUR GWAS, their contribution to the prediction in the EUR population was limited. Our previous explanation of observation, which attributed the limited contribution from non-EUR GWAS to the power saturation of the EUR GWAS seemed to be misleading. We have revised the text in the manuscript accordingly. See Page 7, Line 275-280.

Specific comments:

1. line 162, could the authors provide a short explanation about why “PRS-CS may be less accurate than LDpred when the genetic architecture is highly polygenic (10% causal variants) and the discovery sample size is limited”?

-- Great question. This is something we had observed when we developed the PRS-CS algorithm and benchmarked its performance against LDpred. After updating LDpred to LDpred2, we continued to have this observation although LDpred2 in general improved over LDpred. We suspect different predictive performance of PRS-CS and LDpred2 reflects the strengthens and limitations of the continuous shrinkage prior vs. the spike-and-slab prior used in the two algorithms. Specifically, to ensure that posterior effect size estimates are not inflated, we have imposed a minimal shrinkage to the marginal effect size estimates in PRS-CS, which may induce a stronger-than-optimal regularization when the GWAS sample size is small. In contrast, LDpred2 does not guarantee the boundedness of the posterior effect size estimates, which may better separate signals from noise when the GWAS has limited power, at the expense of the algorithm being sensitive to the imperfectly matched LD reference panel and having convergence issues when the GWAS sample size is large. We have added some discussions on these issues in the revised manuscript. See Page 4-5, Line 168-177.

2. For the analysis of the UKBB dataset, since Neale Lab GWAS was generated based on all the independent EUR population, I am confused why the authors can still have ~14,000 EUR individuals left that are not overlapped with Neale Lab GWAS samples.

-- The Neale Lab GWAS was conducted in White British individuals in the UK Biobank. We defined a larger sample of European ancestry, among which we identified European individuals who are non-British and are unrelated with the Neale Lab GWAS sample as the target dataset for PRS validation and testing. We have clarified this in the revised manuscript. See Page 17, Line 678-680.

3. In figure 2, the authors only provide single-PRS performance based on the EUR population, as the target population is EAS in figure 2, they may also want to provide single-PRS performance based on the EAS population. It is similar for the right panel of figure 4, better to include more methods to make the results more comprehensive.

-- Thanks for the suggestion. We have now included the performance of single-discovery methods trained on the non-EUR GWAS summary statistics in Figure 2.

4. In Table S3, it is not clear why the authors use the raw data for some traits, while use IRNT transformation for others.

-- In the previous version of the BBJ GWAS release, some traits were inverse rank normalized before conducting the GWAS, while other traits were tested for association on their original scale. As the Neale Lab provided both transformed and untransformed GWAS, we tried to match the scale of the phenotype between BBJ and UKBB GWAS. In the revised version of the manuscript, we have updated all BBJ GWAS to the latest version which were conducted on the original scale. In addition, the scale of the phenotype seems to have minimal impact on prediction accuracy. Therefore, we have used GWAS of the raw phenotype in both BBJ and UKBB throughout the revised manuscript.

Reviewer #2:
In this paper, the authors introduced PRS-CSx, an extension of PRS-CS, which shown an improved performance over existing methods when applied to under-represented population (e.g. Africans. The performance of PRS-CSx is promising, and it will lay an important foundation for future development of cross-population PRS algorithms. Some details on the simulation procedure, and phenotype definition for the real phenotype analyses were missing, making it slightly difficult to fully replicate the simulation analyses stated in the current paper (see below). To summarize, PRS-CSx is an important development to the field, but the current paper needed to address a few issues before it can be published. Below are some questions and comments:

1. In the introduction, the authors claimed that "…, PRS alone are already more accurate than combined clinical risk factors currently used for population screening for some diseases…", which seems hard to believe. The same claim was not stated in the two cited papers, instead, those papers suggest that PRS provides additional information on top of existing clinical risk factors e.g. family history.

-- Thanks for the comment. We agree that this statement may be overly strong. We have revised the text to say that "PRS can already provide predictive power above and beyond combined clinical risk factors currently used for population screening for some diseases such as breast cancer in European populations". See Page 2, Line 56-58.

2. How exactly were the phenotype simulated? Usually, phenotypes (Y) were simulated using Y = XB + E where B is the simulated effect size (in this paper, it was simulated using the multivariate normal distribution), E is the non-genetic component, usually assume to follow a normal distribution and X is the genotype. Were the phenotypes simulated using a standardized X? If X is standardized, is it standardized within each sub-population?

-- Apologize that we didn't provide enough details. The phenotype was simulated using $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ in each population, where $\boldsymbol{X}$ was the genotype matrix (not standardized), $\boldsymbol{\beta}$ was the simulated per-allele effect size vector in which causal variants had non-zero effects and the rest of the variants had zero effect sizes, and $\boldsymbol{\epsilon}$ was the simulated non-genetic component drawn from a normal distribution, which was scaled to fix the trait heritability at the desired value. These have now been clarified in the revised manuscript. See Page 16, Line 633-636.

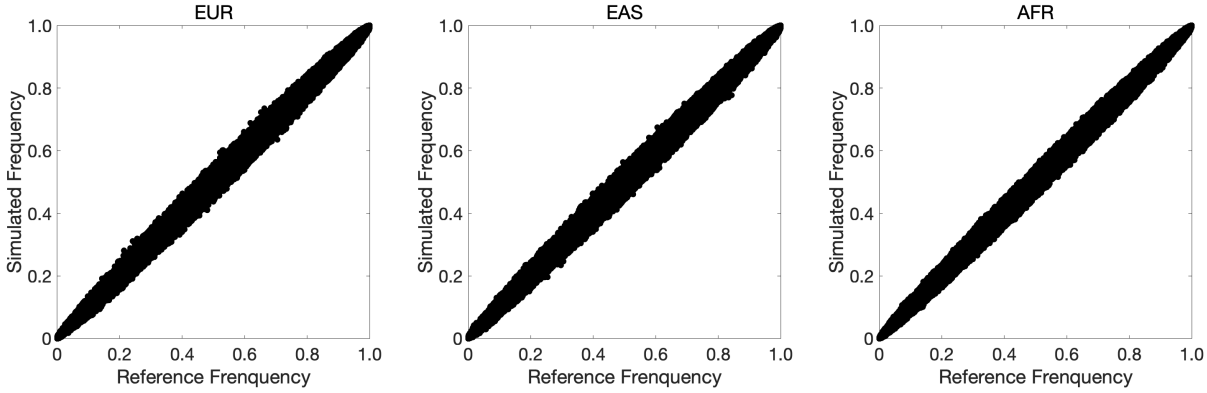
As suggested by Reviewer 3, we have added a new set of simulations where, instead of sampling per-allele SNP effect sizes from a multivariate normal distribution with homogeneous variance across the genome, we assumed that the variance of SNP j in population k is proportional to $[2f_{jk}(1-f_{jk})]^{\alpha}\ell_{jk}^{\alpha}$, where f_{jk} and ℓ_{jk} are the minor allele frequency (MAF) and LD score of SNP j in population k, respectively. $\alpha = 0$ corresponds to the main simulation setting in the manuscript which assumed that per-allele SNP effect sizes are independent of allele frequency and LD patterns. $\alpha = -1$ corresponds to standardized genotype matrices. When $\alpha < 0$, variants with lower MAF and variants located in lower LD regions tend to have larger effects on the trait as observed in recent empirical studies. We used $\alpha = -0.25$ in this set of simulation, which has been empirically estimated to reflect the relationship between effect size and allele frequency and produced approximately a 4-fold difference in the variance of per-allele effect size for both high-frequency vs. low-frequency variants and high-LD vs. low-LD variants included in the simulation. Our results showed that PRS-CSx is robust to the coupling between SNP effect size, allele frequency and LD. See Page 6, Line 238-242; Page 16, Line 647-655; Supplementary Figure 7.

3. The code for interpolation is for interpolation of cM, not for cM/Mb. As the rate (cM/Mb) and cM from the provided file were calculated from different method, it is not clear how the interpolation of rate was done based on the interpolated cM.

-- We linearly interpolated the genetic map (cM) and recombination rate (cM/Mb) using the same method. This has now been clarified in the Methods section. See Page 15, Line 617-621.

4. The 1000G haplotype file provided contains samples from all population. Did the author separated the haplotypes into each super-population before using as an input to HapGen? In similar vein, were there any quality check (e.g. PCA plot) to show that HapGen were simulating samples from different populations?

-- Yes, we separated the 1000 Genomes haplotypes into EUR, EAS and AFR super-populations before using them as the reference panel in HAPGEN2. This has now been clarified in the manuscript. See Page 15, Line 614-617. For each population, we simulated all the samples in a single HAPGEN2 run to avoid potential batch effects in the simulated genotypes. We confirmed that, within each population, the allele frequency and LD patterns of the simulated genotypes were highly similar to those of the 1KG reference panels (e.g., see the frequency plot below), while across populations, the MAF and LD of the simulated genotypes showed substantial differences, suggesting that the simulated data indeed captured cross-population differences in allele frequency and LD patterns.



5. How many SNPs were included in the simulation? Are the causal variants included or excluded in the data used for downstream analyses?

-- We included 1,296,253 HapMap3 SNPs with MAF >0.01 in at least one of the EUR, EAS and AFR populations in the simulation. The exact number of SNPs used in the simulations has been added to the Methods section. See Page 16, Line 626-627. Causal variants were sampled from these HapMap3 SNPs and thus included in downstream analyses. That being said, if a causal SNP was rare (MAF <1%) in one or two of the populations, it would be missing from the LD reference panels for those populations and thus excluded from the construction of Bayesian PRS for those populations. We have clarified these in the revised manuscript. See Page 14, Line 540-544.

6. In the simulation, when assessing the multi-discovery methods, an additional 20k samples were used as a discovery dataset, which makes it difficult to determine whether the improved performance is due to increased sample size or due to the algorithm. It might be worthwhile to also use 100k EUR samples as the discovery dataset for the single discovery methods just to eliminate the sample size different.

-- Thanks for the suggestion. This is a question raised by all reviewers. In the primary simulation, we would like to keep the sample size difference for single-discovery and multi-discovery methods because (i) this design mimics the real-world scenario where each method is applied to the largest GWAS available; (ii) this design shows the advantage of integrating data from multiple populations over single-discovery methods in cross-population prediction. However, we totally agree with the reviewers that it is important to investigate whether the improvement in prediction accuracy can be attributed to increased diversity of the discovery data, or was simply due to sample size increase. We therefore devoted one new set of simulations to address this question where we varied the ratio of the EUR vs. non-EUR GWAS sample sizes with the total sample size kept constant (120K EUR + 0K non-EUR; 100K EUR + 20K non-EUR; 80K EUR + 40K non-EUR; 60K EUR + 60K non-EUR). We observed that the prediction in non-EUR populations benefitted substantially from increasing the proportion of non-EUR training samples, and the coupled shrinkage prior provided consistent gain in prediction accuracy as the power of the non-EUR GWAS varied. See Page 6, Line 225-229; Supplementary Figure 4.

7. On the other hand, when a population specific discovery data is available, it might be beneficial to use it instead of the larger non-population specific discovery, as it might better capture the difference in genetic architecture. It would therefore be interesting to know how the single discovery methods performed when 20k population specific discovery data were used.

-- Thanks for the suggestion. We have now included the performance of single-discovery methods trained on each discovery sample in both simulation studies (Figure 2) and real data analyses (Figures 3 and 4).

8. The authors of LDpred has now published LDpred2, which addresses some issues presented in LDpred. Specifically, in their paper, they have compared the performance of LDpred2, lassosum (a PRS method that utilize penalized regression) and PRS-CS and they have shown that both LDpred2 and lassosum has an improved performance over PRS-CS. Given their superior performance over PRS-CS, it might be interesting to observed how they compared with PRS-CSx when applied to under-represented populations.

-- Thanks for the suggestion and pointing us to the improved version of LDpred. We have updated LDpred results to LDpred2 throughout the manuscript.

9. The procedure used for the multi-discovery method PT-meta and PT-mult is not restrictive to a specific method. The same procedure can be applied to LDpred. Given that PT usually has the worst performance across all methods, it will also be beneficial to include LDpred-meta and LDpred-multi in the comparison.

-- Thanks for the comment, which has been suggested by all reviewers. In the revised manuscript, we have added LDpred2-mult and PRS-CS-mult to all simulations and real data analyses. In particular, a comparison between PRS-CS-mult and PRS-CSx quantifies the benefits of using a coupled shrinkage prior across populations. Our results showed that PRS-CSx consistently outperformed the prediction of PRS-CS-mult in non-European populations, although the amount of improvement varied across traits and was not always substantial. For example, when integrating UKBB and BBJ GWAS and predicting into the EAS population, PRS-CSx had a median relative improvement of 8.3% and 7.1% over PRS-CS-mult in the EAS and AFR populations, respectively (Figure 3). The improvement of PRS-CSx over LDpred-mult was greater, with a median relative increase of 10.5% and 22.2% in the EAS and AFR populations, respectively (Figure 3).

We chose to compare PRS-CSx against LDpred2-mult and PRS-CS-mult rather than LDpred2-meta and PRS-CS-meta because (i) The only modeling difference between PRS-CSx and PRS-CS-mult is the use of a coupled shrinkage prior. Comparing PRS-CSx with the “mult” methods thus provides a direct characterization

of the benefits of the coupled prior. (ii) The LD pattern of a cross-ancestry meta-analyzed GWAS is a mixture of population-specific LD patterns, which is difficult to be appropriately modelled using current methods. Meanwhile, an accurate modeling of the LD patterns is critical to Bayesian polygenic prediction methods. In fact, we have observed convergence issues for LDpred2 when there is mismatch between the discovery GWAS summary statistics and the LD reference panel. We therefore do not recommend LDpred2-meta and PRS-CS-meta from a modeling perspective. The same recommendation was made by other recent studies in the literature (e.g., *medRxiv*: https://doi.org/10.1101/2021.01.19.21249483).

10. For the real data analysis, most of the selected traits are highly correlated (e.g Diastolic blood pressure and Systolic blood pressure; Mean corpuscular hemoglobin and Mean corpuscular hemoglobin concentration etc). As we know performance of PRS software are trait dependent (e.g. LDpred has much better performance for Height), will it be possible to use less correlated traits for the real data analysis (if available)?

-- As we mentioned at the beginning of the response letter, we have now updated all BBJ GWAS to the latest version and expanded the quantitative trait analysis in the UK Biobank (UKBB) and Biobank Japan (BBJ) from 16 traits to 33 traits, and the analysis in the UKBB, BBJ and PAGE study from 7 traits to 14 traits. These traits broadly cover anthropometric, cardiovascular, hematological, kidney, liver and metabolic measures, and are the largest collection with publicly available summary statistics across at least two global populations. We hope that this expanded list of traits has covered a wider range of genetic architectures and provided a more comprehensive comparison of different polygenic prediction methods in real applications.

11. Here, the authors reported the median increase in R2, yet it seems like it should be the median relative increase in R2 (as in some instance the increase is up to 209%)

-- Thanks for the careful reading. Yes, we meant to report the median relative increase in R^2. We have doubled checked the manuscript to make sure that all statistics are accurately reported.

12. Likely typo on line 272 to 275. The author stated that all PRS R2 increased to 0.060, yet figure 4a shows that only PRS-CSx reaches R2 of 0.06.

-- Thanks for pointing this out. We have rephrased this sentence. See Page 9-10, Line 359-362.

13. In the method section, for PT-multi, the authors stated that “The optimal p-value threshold for each discovery sample and the weights for the linear combination …”, what is the weight?

-- We referred to the coefficients for the linear combination of multiple PRS, which has now been clarified. See Page 15, Line 600-609.

14. What is the number of SNPs remained after quality controls were performed on the UK Biobank? Were the imputed data used for the PRS analysis? If imputed data were used, what were the thresholds used to transform the dosage information into hard coded genotypes?

-- Yes, imputed UK Biobank data was used. The final target dataset included 12,886,200, 8,593,932, 6,506,126, 8,211,053 and 8,032,121 variants for the AFR, AMR, EAS, EUR and SAS populations, respectively. We used the default parameters in PLINK 2.0 to convert dosage information into hard coded genotypes (i.e., dosage was rounded to the nearest hardcall when the distance was no great than 0.1; otherwise a missing hardcall was saved). These have now been reported in the manuscript. See Page 17, Line 680-686.

Reviewer #3:
The manuscript by Ryan et al presents an approach for PRS that attains improved accuracy in cross-population settings. The approach is a marginal extension of the recently proposed PRS-CS method from the same authors from one to multiple populations. The approach is compared with other recently proposed methods for PRS and shows improved prediction (as measured by R2) in simulations and real data from UK Biobank.

1. I found it unclear how is the current method PRC-CSx different from PRC-CS. The math on page 12 is very brief and does not really explain how are the causal effects coupled across populations. From the brief second equation in Methods, where I assume j stands for a SNP, the causal effects across populations are drawn from a gaussian with population specific variance parameter that is scaled by \psi_j variable that is similar across populations. What are different choices of \psi_j implying on the cross-population genetic correlation? Second, how robust is the inference across the prior parameters (a=1, b=1/2)? Some simulations or analytical derivations on impact of this prior on the cross-population genetic correlation in causal effects would be very useful.

-- Thanks for the comment. In the revised version of the manuscript, we have expanded the Methods section to provide a more detailed description of the PRS-CSx model. In particular, we have analytically derived the posterior mean of the SNP effect sizes, which was then used to explain how genetic effects are coupled across populations. See Page 14, Line 528-536. Additional information on the PRS-CSx model and posterior inference has been included in the Supplementary Methods. We hope that the method has now been more clearly described.

The hyper-parameters of the coupled continuous shrinkage prior (a=1, b=1/2) were inherited from PRS-CS. Parameter a controls the shape of the continuous shrinkage prior around zero, while b affects the tails of the prior distribution. A smaller value of a places more probability mass near zero, and thus imposing stronger shrinkage on the SNP effect sizes; a smaller value of b produces heavier tails of the prior distribution, and thus can better accommodate large genetic effects. When developing PRS-CS, we evaluated the impact of the two hyper-parameters on prediction accuracy using a small two-dimensional grid and concluded that the predictive performance is robust to the selection of a (a=1/2, a=1 or a=3/2) as long as substantial probability mass of the prior distribution is placed at the original, while it is important to set b to a small value such that the prior has heavy, Cauchy-like tails in order not to over-shrink truly non-zero effects. In the revised manuscript, we added a new set of simulations to re-evaluate the robustness of the PRS-CSx algorithm with respect to these two hyper-parameters. Specifically, we assessed several combinations of the two parameters (a=1/2, b=1/2; a=1, b=1/2; a=3/2, b=1/2; a=1, b=1), and had the same observations as before: the predictive performance was robust to the selection of a but setting b to larger values substantially reduced prediction accuracy. See Page 6, Line 242-246; Supplementary Table 9. We therefore fixed the two parameters to a=1 and b=1/2, which is consistent with the default parameter setting of PRS-CS. In practice, we do not recommend tuning these two hyper-parameters when using PRS-CS or PRS-CSx.

2. Figure 2. The authors compare multi-population methods vs single-population by adding 20k new samples from the target population only for multi-population methods. That conflates the gain in accuracy due to boost in sample size, or better modeling of cross-population causal effect. I would encourage the authors a plotting style for which the sample size is fixed across the different comparator methods. Similarly for Figure 3.

-- Thanks for the suggestion. This is a question raised by all reviewers. In the primary simulation, we would like to keep the sample size difference for single-discovery and multi-discovery methods because (i) this design

mimics the real-world scenario where each method is applied to the largest GWAS available; (ii) this design shows the advantage of integrating data from multiple populations over single-discovery methods in cross-population prediction.

However, we totally agree with the reviewers that it is important to investigate whether the improvement in prediction accuracy can be attributed to increased diversity of the discovery data and the joint modeling of genetic effects across populations, or was simply due to sample size increase. We therefore added two analyses to address this question. First, we added one new set of simulations where we varied the ratio of the EUR vs. non-EUR GWAS sample sizes with the total sample size kept constant (120K EUR + 0K non-EUR; 100K EUR + 20K non-EUR; 80K EUR + 40K non-EUR; 60K EUR + 60K non-EUR). We observed that the prediction in non-EUR populations benefitted substantially from increasing the proportion of non-EUR training samples, and the coupled shrinkage prior provided consistent gain in prediction accuracy as the power of the non-EUR GWAS varied. See Page 6, Line 225-229; Supplementary Figure 4. Second, we have added LDpred2-mult and PRS-CS-mult to all simulations and real data analyses. In particular, the only modeling difference between PRS-CSx and PRS-CS-mult is the use of a coupled shrinkage prior, and thus comparing PRS-CSx with PRS-CS-mult directly characterizes the benefits of jointly modeling SNP effect sizes across populations. These analyses showed that increasing diversity of the discovery GWAS and coupling genetic effects across populations indeed improved prediction in non-EUR populations. We hope that these additions have sufficiently addressed the concern.

3. Results in real data show the proposed method outperforms others. First, it is hard to compare methods that use variable sample sizes (see comment above). Second, the authors focus on relative R2 by normalizing to prediction in EUR which mask out the main prediction r2 that are rather small and likely not useful in practice: e.g., for AFR using UKBB+BBJ, the average absolute R2 is 2% PRS-CSx, 1% (PT-meta) and 1.3% (PT-multi) (Table S4).

-- Respectfully, we argue that, in real data analysis, the comparison across different polygenic prediction methods was fair in the sense that each method was applied to the largest GWAS available. This is also the practice being used in other cross-ancestry PRS manuscripts (e.g., Tables 3-5 in PMID 29110330; *medRxiv*: https://doi.org/10.1101/2021.01.19.21249483). Since we don't have access to individual-level data for many biobanks (e.g., Biobank Japan and the PAGE study) and rely on publicly available GWAS summary statistics, it is not practical to rerun all the GWAS and make the total discovery sample size equal across methods. However, as we mentioned, all the multi-discovery methods (PT-meta, PT-mult, LDpred2-mult, PRS-CS-mult, PRS-CSx) had the same discovery sample size, and a comparison between PRS-CSx and PRS-CS-mult directly quantifies the benefits of the coupled continuous shrinkage prior.

We agree that for some traits, after combining discovery GWAS from multiple populations, the prediction accuracy remained low especially in non-EUR populations and may not be useful in practice. We think this primarily reflects the relatively low heritability and the lack of powerful non-EUR GWAS for many traits, rather than a weakness of the method. The low absolute prediction accuracy in non-European ancestries reflects the long-time Eurocentric biases in GWAS that requires systematic efforts to overcome. Our methodological innovation is a critical first step towards addressing this challenge by providing a tool that maximizes the utility of the non-European data, which encourages more genetics studies in the non-European populations. We have added some discussions surrounding these points to the manuscript. See Page 12, Line 449-458.

We chose to present relative prediction accuracy in the main text because we examined a wide range of traits with different heritability, genetic architectures and prediction accuracy. Reporting the median absolute R^2 may be misleading when comparing methods. For example, it is theoretically possible that Method A improves

prediction for all but one trait over Method B but has a lower median absolute R^2. As the primary goal of this manuscript is to develop a cross-population polygenic prediction method and benchmark its performance against alternative methods, we believe that it is more appropriate and informative to present the relative R^2 in the main figure. We have included all absolute R^2 estimates in Supplementary Tables.

4. The simulations could be improved to be more realistic. First some correlation in the non-genetic component could be considered. Most importantly it would be useful to assess the accuracy under more realistic couplings of causal effects with MAF and LD (i.e. by using the baseline-LD model). Third, how do the authors deal with variants that are monomorphic in one of the populations?

-- Thanks for the suggestions. First, since we focused on cross-population prediction in this work, where discovery GWAS samples were non-overlapping and had different genetic ancestry, we think it is reasonable to assume that non-genetic (environmental) components are independent across samples, which was also the practice of other cross-ancestry PRS studies (see e.g., PMID 29110330; *medRxiv*: https://doi.org/10.1101/2021.01.19.21249483). Correlation in the non-genetic component may be relevant for within-population cross-trait prediction, where the training samples may be overlapping for different traits. As we discussed in the manuscript, the PRS-CSx framework may be extended to integrate genetically correlated traits to inform and improve cross-trait prediction (see Page 12, Line 445-447), but we would love to save this as a future direction.

To address the second suggestion, in the revised manuscript, we have added a new set of simulations where instead of sampling per-allele SNP effect sizes from a multivariate normal distribution with homogeneous variance across the genome, we assumed that the variance of SNP j in population k is proportional to $[2f_{jk}(1-f_{jk})]^{\alpha}\ell_{jk}^{\alpha}$, where f_{jk} and ℓ_{jk} are the minor allele frequency (MAF) and LD score of SNP j in population k, respectively. $\alpha = 0$ corresponds to the main simulation setting in the manuscript which assumed that per-allele SNP effect sizes are independent of allele frequency and LD patterns. $\alpha = -1$ corresponds to standardized genotype matrices. When $\alpha < 0$, variants with lower MAF and variants located in lower LD regions tend to have larger effects on the trait as observed in recent empirical studies. We used $\alpha = -0.25$ in this set of simulation, which has been empirically estimated to reflect the relationship between effect size and allele frequency and produced approximately a 4-fold difference in the variance of per-allele effect size for both high-frequency vs. low-frequency variants and high-LD vs. low-LD variants included in the simulation. Our results showed that PRS-CSx is robust to the coupling between SNP effect size, allele frequency and LD. See Page 6, Line 238-242; Page 16, Line 647-655; Supplementary Figure 7. The expanded real data analysis in the biobanks also showed the robustness of PRS-CSx to varying genetic architectures.

Third, monomorphic or rare variants not present in the population-specific LD reference panel of population A are not included in the construction of PRS for population A. If a SNP is present in population A but is monomorphic or rare in other populations, its effect size is not coupled in posterior inference but the SNP is included in the PRS of population A such that population-specific associations can be captured. We have clarified these in the Methods section of the revised manuscript. See Page 14, Line 540-544.

5. Some standard MCMC metrics of performance would be useful. How many chains are employed and what is the autocorrelation across samples?

-- Great question. All the prediction accuracy reported in the manuscript was based on one Markov chain. The convergence of MCMC samplers employed by Bayesian polygenic prediction methods is often overlooked in the literature, likely because (i) the focus of polygenic prediction is to aggregate genetic effects across the genome into a single score, rather than making inference of the genetic effects of individual variants; and (ii)

the convergence of the MCMC algorithm is typically monitored by saving the Markov chain for all model parameters and assessing convergence either visually or by diagnostic metrics. However, in polygenic prediction the sheer size of the model parameters (>1 million for each population) makes traditional model diagnostic methods difficult to apply.

To assess the overall convergence of the Gibbs sampler used in PRS-CSx, for each trait, we selected a few SNPs where we monitored the convergence of their posterior effect size estimates. Some of the SNPs had strong associations with the trait in multiple populations, while some of the SNPs were null across populations. We ran the PRS-CSx model three times using different random seeds, and assessed the convergence using the Gelman-Rubin convergence diagnostic for multiple chains. All reduction factors across the SNPs we examined were smaller than 1.05, indicating convergence. As an example, Supplementary Fig. 8 shows the trace plots and autocorrelation functions (ACFs) for the posterior effects of rs7412 on low-density lipoprotein cholesterol (LDL-C) when integrating UKBB, BBJ and PAGE GWAS summary statistics using PRS-CSx. This SNP, located within the *APOE* locus on chromosome 19, had extremely strong marginal associations with LDL-C across the three populations (all *P*-values <1E-200). Trace plots and ACFs indicated that the Gibbs sampler achieved reasonable convergence and mixing. We also note that PRS-CSx guarantees that the posterior effect size estimates are bounded by the marginal effects (see the posterior mean derivation on Page 14, Line 532), while other Bayesian polygenic prediction methods that rely on discrete mixture priors do not have this guarantee and may produce wildly inflated effect size estimates. We have added some of these text as well as a discussion on future directions to the manuscript. See Page 11, Line 413-430.

Decision Letter, first revision:

15th Oct 2021

Dear Dr Huang,

Your Article, "Improving Polygenic Prediction in Ancestrally Diverse Populations" has now been seen by 3 referees. You will see from their comments below that while they find that the manuscript has improved, some final points are raised. We are very interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision.

As you will see, all reviewers state that the manuscript has improved and appreciate the extensive additional analyses added. Reviewer #1 continues to think that the marginal improvement of the prediction accuracy will limit practical utility, and has some specific comments. Reviewer #2 has some remaining points that should be addressed. Reviewer #3 raises a few minor points that can be addressed with textual changes.
As the manuscript has been significantly improved and there are no major outstanding technical issues, we are happy to move forward with this. We ask that you address the points raised by all 3 reviewers, and it would be helpful if you could clarify or make a better case for the utility of the approach.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the

manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Thank you very much.

All the best,

Catherine

Catherine Potenski, PhD
Chief Editor
Nature Genetics
1 NY Plaza, 47th Fl.
New York, NY 10004
catherine.potenski@us.nature.com
https://orcid.org/0000-0002-4843-7071

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The manuscript has certainly been improved with the addition of (1) more comprehensive simulation analysis, (2) the comparison between PRS-CSx and LDPred2-multi/PRS-CS-multi, and (3) analysis including the Taiwan data and LDPred2. Thank the authors for making all these efforts!

I overall remained concerned about the practical usefulness of the method, given the marginal improvement of the prediction accuracy over the LDPred2-multi and PRS-CS-multi. Below are more specific comments:

1. From Table S11, among the 33 quantitative traits, 7 traits are better predicted by PRS-CS-multi in AFR, 9 traits are better predicted by LDPred2-multi in AMR, 6 traits are better predicted by LDPred2-multi in EAS, and 15 traits are also better predicted by LDPred2-multi in EUR when compared to PRS-CSx, which challenges the robustness of the superiority of PRS-CSx.
2. The authors argue that the marginal improvement is probably due to the lack of powerful non-EUR GWAS, but their results have shown that for those well-powered EUR GWAS, adding other non-EUR GWAS could not have further significant improvement, which indicates that when the GWAS is already well-powered, the method may also not have much gain over the existing method like LDPred2-multi and PRS-CS-multi. The authors may want to better clarify in which scenario this method can bring significant additional gains.
3. Figure 4: from (a), the PRS-CSX only has a tiny improvement over the LDPred2-multi and PRS-CS-multi when applied to schizophrenia. Especially for the Japanese population, LDPred2-multi is even significantly better than PRS-CS-multi, but the authors didn't have any discussions on that. And I'm confused why the authors choose to show the comparison between PRS-CSx and LDPred2 in figure 4(b), instead of comparison between PRS-CSx and LDPred2-multi.
4. The authors mention the MCMC convergence issue when trying to create PRS-CS-meta or LDPred2-meta, it will be appreciated if the authors can provide the corresponding MCMC diagnostic plots and corresponding prediction accuracy, which should be relatively small if the MCMC convergence issue is prominent. I am curious here because, in the recent paper https://doi.org/10.1016/j.ajhg.2021.03.002, they have tried the reference panels from both populations, and made a persuasive comparison between their method XPASS and Ldpred-meta.

Reviewer #2:
Remarks to the Author:
In this paper, the authors have introduced an extension to PRS-CS, PRS-CSx, which has improved performance when applied to under-represented population. In this update, the authors have done a thorough job in addressing most of the concerns, providing additional simulations and analyses, and have provided much needed details in the method sections. A methods for cross-population PRS analyses is much needed in the field and the development of PRS-CSx is definitely a welcoming sight. I have but a few questions remaining:

1. While I agree that the relative R2 is a good metric of relative performance of different software, given that the PRS R2 can often be small, especially in non-European samples (for example, the

median R2 for AFR samples across all methods is around 0.0097 vs 0.0509 in EUR), it might still be worthwhile to at least show the difference of different population vs EUR (e.g. Performance of different method in different population vs performance of PRS-CS in EUR). Using this metric, we can see that while PRS-CSx outperforms PRS-CS using the EUR GWAS in all population, it still underperformed when compared to PRS-CS in EUR samples using EUR GWAS (except for AMR, which is interesting)

2. Throughout the paper, the authors report the relative increase of performance, however, in figure 3, the ratio of performance were reported instead (R2 / R2PRS-CS (UKBB) instead of (R2 - R2PRS-CS (UKBB)) / R2PRS-CS (UKBB)). It might be best to keep the reported metric consistent.

3. On page 6 line 199-200, the authors stated that "However, when predicting into non-EUR populations, Bayesian multi-discovery methods demonstrated a clear advantage over single discovery methods." Based on the results, it seems like all multi-discovery methods, including PT-meta and PT-mult also out-perform the single discovery methods most of the time.

4. If I understand correctly, to use PRS-CSx, we will first perform the PRS-CSx analysis to obtain the adjusted summary statistics for k populations across a few \phi. TO obtain the PRS-CSx PRS, we then need to fit the k PRS in a linear regression model (or logistic regression for binary trait), and then obtain the individual coefficients. The "final" PRS is then calculated by applying the coefficient from the linear regression to the k PRS calculated for the validation data set. If this is the case, will it be possible for us to use PRS-CSx in a relatively small cohort where split in half analyses might not be viable due to power, or if we would like to obtain PRS for all samples within our cohort? This might be a common scenario as sample size for the under-represented population are usually small.

5. Another question I have is regarding non-overlapping SNPs. On page 14 line 538-542, it stated that monomorphic or rare variants do not present in the population specific LD panel for population A are not included in the construction of PRS for population A, but if the variants were missing in other population but not in population A, it will be included in the PRS calculation but not coupled in the posterior inference. Does that mean that for any SNPs that were only presented in one of the populations, the effect size of those SNPs will be applied as is without any shrinkage? To help me better understand this, imagine an extreme scenario where we have two populations A and B and the genotyping chips of these two populations does not share any SNPs. In this unlikely scenario, is it safe to assume that PRS-CSx will reduce into PRS-CS, where each of the populations were essentially analyzed separately and have their own adjustment?

Reviewer #3:
Remarks to the Author:
The revised manuscript by Ryan et al present new simulations and analyses that clarify and improve the manuscript. I thank the authors for the detailed comments with my remaining comments being on text clarifications.

1. I could not find the derivation for the statement that under the proposed model the posterior mean effects is (D_k + T^-1)^2\beta_hat_k. In particular, the equation describing \beta_jk from main text is different from the supplementary. It appears \phi could be propagated through the gamma distributions to reach similar equations, but why the different equations? Also, no derivation is provided for the MCMC equations being correct to sample from the posterior of \beta as defined in the

model. I encourage the authors to tighten up their presentation of the mathematical aspects of their approach such that the reader can replicate/understand the mathematical details of their model.

2. I commend the authors for the greatly expanded simulations exploring various parameters and impact of architecture on their results (Supp Figs 1-7).

3. The authors continue to present their results in main text conflating sample size vs PRS method in assessing improvements in PRS performance. While it is ultimately the authors choice on how to present their main results, I continue to find this style of presentation likely to lead to misleading interpretations.

Author Rebuttal, first revision:

NG-A56583

Improving Polygenic Prediction in Ancestrally Diverse Populations

Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Stanley Global Asia Initiatives, Lin He, Akira Sawa, Alicia R. Martin, Shengying Qin, Hailiang Huang, Tian Ge

We thank all the reviewers for their critical and constructive comments which have led to this improved manuscript. In this revision, we have further addressed the remaining concerns from the reviewers, with the point-by-point response provided below and the changes to the manuscript highlighted in red.

Reviewer #1

Remarks to the Author:

The manuscript has certainly been improved with the addition of (1) more comprehensive simulation analysis, (2) the comparison between PRS-CSx and LDPred2-multi/PRS-CS-multi, and (3) analysis including the Taiwan data and LDPred2. Thank the authors for making all these efforts!

I overall remained concerned about the practical usefulness of the method, given the marginal improvement of the prediction accuracy over the LDPred2-multi and PRS-CS-multi. Below are more specific comments:

1. From Table S11, among the 33 quantitative traits, 7 traits are better predicted by PRS-CS-multi in AFR, 9 traits are better predicted by LDPred2-multi in AMR, 6 traits are better predicted by LDPred2-multi in EAS, and 15 traits are also better predicted by LDPred2-multi in EUR when compared to PRS-CSx, which challenges the robustness of the superiority of PRS-CSx.

We thank the reviewer for raising this critical comment. We do agree with the reviewer that, when predicting into the European population, adding a smaller non-EUR GWAS to a well-powered EUR GWAS provides limited gain in the prediction accuracy, and the relative improvement of PRS-CSx over LDpred2-mult and PRS-CS-mult is marginal. This observation was consistent across our simulation studies and biobank analyses. We have further clarified and discussed this in the revised manuscript in response to this and the next comment (Page 11, Line 409-419).

However, when predicting into non-European populations, PRS-CSx outperformed LDpred2-mult and PRS-CS-mult by improving prediction accuracy for the majority of the traits we examined (PRS-CSx vs. LDpred2-mult: AFR 28/33, AMR 24/33, EAS 27/33, SAS 32/33; PRS-CSx vs. PRS-CS-mult: AFR 27/33, AMR 25/33, EAS 32/33, SAS 32/33). Several traits for which PRS-CSx showed decreased prediction accuracy were not meaningfully predicted by any of the PRS methods (all R^2 <0.5%; e.g., the prediction of AST, GLC and MCHC in the AFR population), and thus the ranking of methods for the prediction of these traits can be unreliable. Overall, the magnitude of improvement in prediction accuracy from PRS-CSx relative to LDpred2-mult and PRS-CS-mult was much larger than the magnitude of decrease in predictive performance. We now formally test and justify this statement by reporting the P-values of the two-sided Wilcoxon signed-rank test in the revised manuscript (PRS-CSx vs. LDpred2-mult: AFR 2.38E-5, AMR 4.25E-3, EAS 3.90E-4, SAS 6.48E-7; PRS-CSx vs. PRS-CS-mult: AFR 2.99E-5, AMR 5.27E-4, EAS 2.84E-6, SAS 9.59E-7; see Page 7, Line 293-294; Page 8, Line 300). All tests were significant and survived multiple testing corrections, suggesting statistically significant improvement of PRS-CSx over LDpred2-mult and PRS-CS-mult in non-European populations.

We note that the process of choosing the best method from a range of PRS methods (e.g., LDpred2-mult and PRS-CS-mult) to compare with PRS-CSx also unfairly puts PRS-CSx at a disadvantage because this practice creates an “ensemble” of methods that naturally outperforms any single method (akin to the concept of ensemble learning). We argue that, PRS-CSx is to date the only principled Bayesian method explicitly designed for multi-ancestry PRS analysis, and when choosing a single method for PRS analysis across populations *a prior*, PRS-CSx is clearly the method of choice because it outperformed any other method for the majority of the traits examined; and for traits where the predictive performance decreased, the loss of prediction accuracy was often minimal.

Lastly, with respect, we note that LDpred2-mult and PRS-CS-mult were created as “intermediate methods” between existing single-discovery PRS methods and PRS-CSx in this work to disentangle the various contributing factors to the increase of the predictive performance from PRS-CSx. While they were certainly helpful for this purpose -- we thank the

reviewer for suggesting this -- the comparison of PRS-CSx with LDpred2-mult and PRS-CS-mult is a bit unfair, as they are not published methods *per se* and have not been used in any prior publication to the best of our knowledge. We have now clarified this in the revised manuscript (Page 3, Line 127-132; Page 4, Figure 1 caption).

2. The authors argue that the marginal improvement is probably due to the lack of powerful non-EUR GWAS, but their results have shown that for those well-powered EUR GWAS, adding other non-EUR GWAS could not have further significant improvement, which indicates that when the GWAS is already well-powered, the method may also not have much gain over the existing method like LDPred2-multi and PRS-CS-multi. The authors may want to better clarify in which scenario this method can bring significant additional gains.

Thanks for raising this important point. PRS-CSx provides limited gain in prediction power when a well-powered GWAS in the target population already exists and GWAS from other populations have smaller sample sizes and lower statistical power. Due to the current Eurocentric bias of genomic research, this happens almost exclusively for predictions in the EUR population. Intuitively, when a well-powered EUR GWAS is available, adding a smaller non-EUR GWAS provides limited benefits for the prediction in EUR individuals. This is consistent with our simulation studies and biobank analyses.

In contrast, PRS-CSx can bring gains in prediction power (although the amount of improvement varies across traits as discussed in response to comment #1 and in the manuscript; see e.g., Page 11, Line 420-425) when the GWAS in the target population has lower statistical power, while well-powered GWAS from other populations are available. This often happens when predicting into a non-EUR population, where ancestry-matched GWAS have limited sample sizes but large-scale EUR GWAS already exist. By integrating EUR and non-EUR GWAS, PRS-CSx can significantly improve the prediction accuracy in non-EUR populations, which alleviates the imminent challenge of polygenic prediction in under-represented populations. In the revised manuscript, we have clarified the scenarios in which PRS-CSx are expected to provide larger power gains (Page 11, Line 409-419).

3. Figure 4: from (a), the PRS-CSX only has a tiny improvement over the LDPred2-multi and PRS-CS-multi when applied to schizophrenia. Especially for the Japanese population, LDPred2-multi is even significantly better than PRS-CS-multi, but the authors didn't have any discussions on that. And I'm confused why the authors choose to show the comparison between PRS-CSx and LDPred2 in figure 4(b), instead of comparison between PRS-CSx and LDPred2-multi.

We agree with the reviewer that PRS-CSx provided relatively small improvement in prediction accuracy compared with LDpred2-mult and PRS-CS-mult when applied to schizophrenia. We

chose schizophrenia as an example of a dichotomous trait out of data availability: at the time we had access to individual-level genotype data for multiple schizophrenia cohorts, which enables a relatively comprehensive assessment of the performance of different PRS construction methods. Being one of the most polygenic disorders, schizophrenia actually represents an application scenario that is not favorable to the PRS-CS-type of methods. We observed in both simulations and the biobank analyses that (i) the benefits of the coupled prior decreased with the increasing polygenicity of the genetic architecture (see e.g., Page 6, Line 217-223; Supplementary Figure 1); (ii) LDpred2 was more accurate than PRS-CS when the discovery sample size was limited (relative to the polygenicity of the trait), and a larger discovery sample size is needed for PRS-CS to outperform LDpred2 for highly polygenic traits (see e.g., Page 6, Line 228-233; Supplementary Figure 3). With further methodological discussions in the manuscript (Page 4-5, Line 174-184), this reflects the strengths and limitations of the continuous shrinkage prior vs. the spike-and-slab prior used in PRS-CS and LDpred2, respectively. For schizophrenia this is evident in the comparison of single-discovery PRS methods: LDpred2 outperformed PRS-CS in 5 out of 6 of the cohorts. Conversely, when integrating EUR and EAS schizophrenia GWAS, PRS-CSx outperformed LDpred2-mult in 5 out of 6 of the cohorts with a median increase in prediction accuracy of 8.7%. This multi-ancestry finding is therefore encouraging and demonstrates the benefits of the coupled prior in practice even for highly polygenic traits. As the GWAS sample size continues to grow in non-European populations, PRS-CSx is expected to provide larger and more robust improvement over LDpred2-mult and PRS-CS-mult for these highly polygenic traits, as shown in the simulation studies. We have added a brief discussion of these points to the revised manuscript (Page 10-11, Line 383-387).

Admittedly, PRS-CSx and LDpred2-mult had largely overlapping confidence intervals in this analysis as shown in Supplementary Table 18. We chose to present the results of LDpred2 rather than LDpred2-mult because, as we discussed in response to comment #1, LDpred2-mult is not a published method nor has been used in any prior publication. It was created in this work to delineate various contributing factors to the predictive performance of PRS-CSx. While it served its designated purpose well, in Figure 4b we prefer to compare PRS-CSx with existing methods that are used in the field to advocate for using a more sophisticated PRS method to make the best use of the non-European samples. To date, the prevailing practice has always been using single-discovery methods or naive multi-discovery methods such as PT-meta and PT-mult. We chose LDpred2 in Figure 4b because it was the best-performing single-discovery method and outperformed published multi-discovery methods including PT-meta and PT-mult.

4. The authors mention the MCMC convergence issue when trying to create PRS-CS-meta or LDPred2-meta, it will be appreciated if the authors can provide the corresponding MCMC diagnostic plots and corresponding prediction accuracy, which should be relatively small if the MCMC convergence issue is prominent. I am curious here because, in the recent paper

https://doi.org/10.1016/j.ajhg.2021.03.002, they have tried the reference panels from both populations, and made a persuasive comparison between their method XPASS and Ldpred-meta.

Thanks for the suggestion and apologies for not providing a more detailed explanation of the exclusion of LDpred2-meta and PRS-CS-meta from the current work in the previous response letter. We decided not to include LDpred2-meta and PRS-CS-meta in this manuscript early in the project for the following three reasons based on our simulation results in selected settings.

First, the mismatch between the LD pattern of a cross-ancestry meta-analyzed GWAS (which is a mixture of population-specific LD) and the reference LD leads to convergence issues for the LDpred2 algorithm. For example, in our primary simulation, when applying LDpred2 to the EUR+EAS meta-GWAS, 73% of the MCMC runs using the EAS reference panel and 63% of the runs using the EUR reference panel had posterior SNP effect size estimates diverged to infinity (individualized polygenic scores >1E+30). Similarly, when applying LDpred2 to the EUR+AFR meta-GWAS, 94% of the runs using the AFR reference panel and 62% of the runs using the EUR reference panel had diverged posterior effect size estimates. Since LDpred2 screens a wide range of hyper-parameter values from which the optimal value is selected, it usually still produces reasonable predictions in the testing sample even if a significant proportion of the tested hyper-parameter values leads to divergence of the MCMC algorithm, which masks the convergence issue. For example, in our primary simulation, when the target population was EAS, the prediction accuracy of PRS-CSx vs. LDpred2-meta was 0.168 vs. 0.141; when the target population was AFR, the prediction accuracy of PRS-CSx vs. LDpred2-meta was 0.117 vs. 0.097. However, as the GWAS sample size increases and the model fitting algorithm becomes more sensitive to LD mismatch, the convergence issue can be worse which further limits the prediction accuracy. For example, in our simulation setting that combined 300K EUR with 60K EAS or AFR samples, when applying LDpred2 to the EUR+EAS meta-GWAS, 96% of the MCMC runs using the EAS reference panel and 89% of the runs using the EUR reference panel diverged; when applying LDpred2 to the EUR+AFR meta-GWAS, all of the runs using the AFR reference panel and 88% of the runs using the EUR reference panel diverged. In this scenario, PRS-CSx had substantially better prediction accuracy than LDpred2-meta in both EAS and AFR populations: 0.274 vs. 0.169 and 0.223 vs. 0.121, respectively. Our observations are consistent with recent studies which found that PRS-CS was more robust to LD mismatch than other Bayesian PRS methods (medRxiv preprint: https://doi.org/10.1101/2021.01.19.21249483).

Second, the predictive performance of LDpred2-meta and PRS-CS-meta heavily depends on whether the assumption of the fixed-effect meta-analysis (i.e., consistent SNP effects across populations) is accurate. While the prediction accuracy of LDpred2-meta and PRS-CS-meta may approach PRS-CSx when the genetic architecture of the trait is highly consistent across

populations, the performance drops dramatically when the assumption is violated. For example, in our simulation setting where the cross-population genetic correlation was set to 0.4, the prediction accuracy of PRS-CSx vs. LDpred2-meta vs. PRS-CS-meta was 0.128 vs. 0.085 vs. 0.087 when the target population was EAS, and 0.085 vs. 0.058 vs.0.059 when the target population was AFR. Therefore, unlike PRS-CSx or the "mult" methods that can be adaptive to a wide range of cross-population genetic architectures, the "meta" methods are competitive in much narrower application scenarios.

Third, running Bayesian polygenic prediction methods, and LDpred2 in particular, is computationally expensive. Each LDpred2 run requires ~20 hours of computational time, ~50 Gb of memory, and writes ~50 Gb of temporary data to the hard disc. Producing full results for LDpred2-meta and PRS-CS-meta across all simulation settings and real data applications will add weeks if not months of computational time, as well as tremendous computational cost, to our already intensive simulations and biobank analyses.

For these reasons, we did not include LDpred2-meta and PRS-CS-meta in this manuscript. That said, we do realize that, although LDpred2-meta and PRS-CS-meta are suboptimal from a modeling perspective and may produce less accurate predictions than PRS-CSx or "mult" methods, many existing studies have only released cross-population meta-GWAS, in which case LDpred2-meta and PRS-CS-meta remain useful approaches in practice. We have added discussion of these points to the revised manuscript, and have advocated the release of ancestry-specific summary statistics from multi-ancestry genomic studies to enable flexible and accurate cross-population polygenic modeling and prediction (Page 11-12, Line 431-443).

Reviewer #2

Remarks to the Author:

In this paper, the authors have introduced an extension to PRS-CS, PRS-CSx, which has improved performance when applied to under-represented populations. In this update, the authors have done a thorough job in addressing most of the concerns, providing additional simulations and analyses, and have provided much needed details in the method sections. A method for cross-population PRS analyses is much needed in the field and the development of PRS-CSx is definitely a welcoming sight. I have but a few questions remaining:

1. While I agree that the relative R2 is a good metric of relative performance of different software, given that the PRS R2 can often be small, especially in non-European samples (for example, the median R2 for AFR samples across all methods is around 0.0097 vs 0.0509 in EUR), it might still be worthwhile to at least show the difference of different population vs EUR (e.g. Performance of different methods in different populations vs performance of PRS-CS in

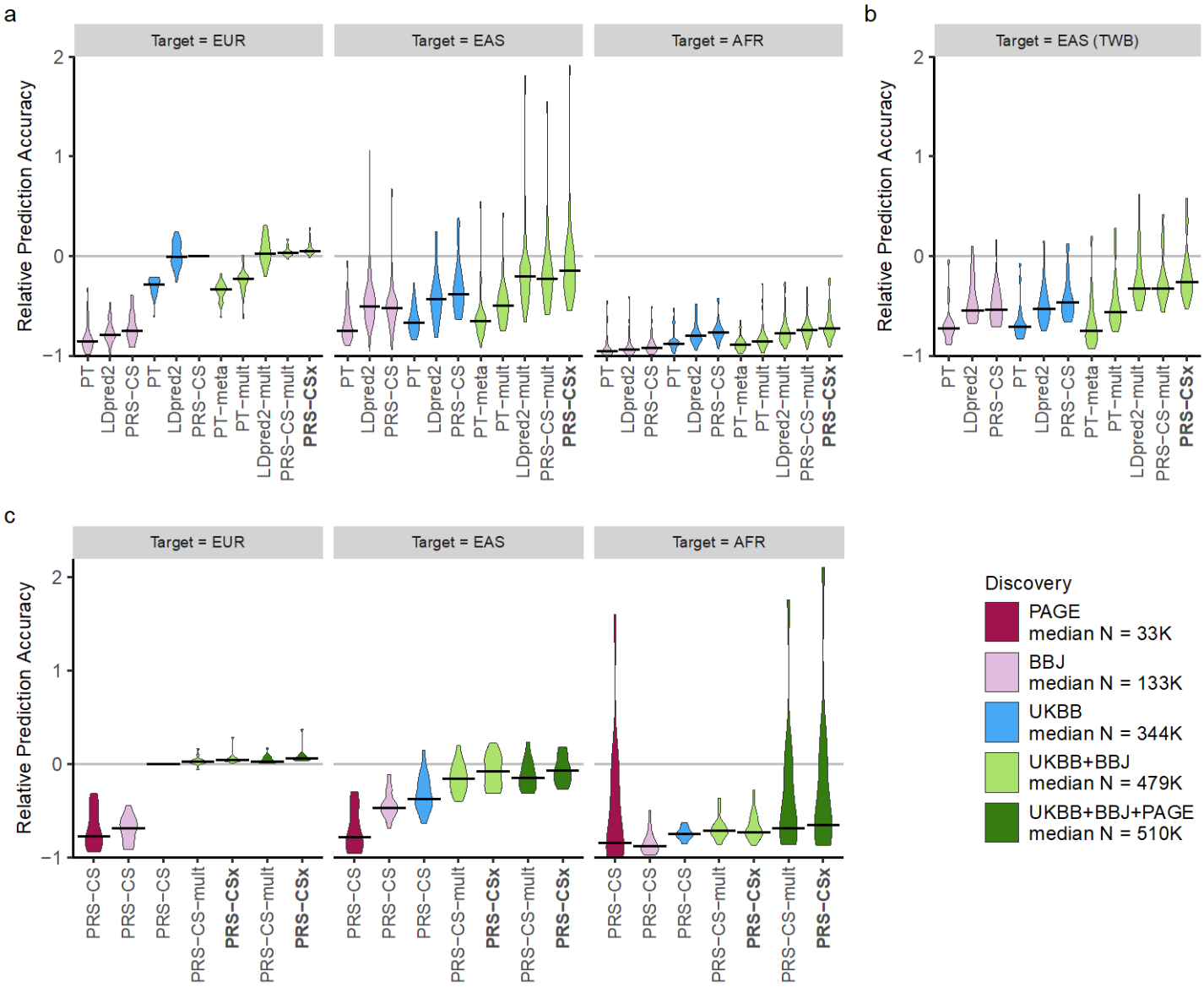
EUR). Using this metric, we can see that while PRS-CSx outperforms PRS-CS using the EUR GWAS in all populations, it still underperformed when compared to PRS-CS in EUR samples using EUR GWAS (except for AMR, which is interesting).

Thanks for the suggestion. We fully agree with the reviewer that this is an important point. We have added Supplementary Figure 8 in the revised manuscript (also pasted below for convenience), which shows the predictive performance of different PRS methods with respect to the prediction accuracy of PRS-CS in the EUR target population trained on the UKBB GWAS. We have additionally noted in the following revised text that while PRS-CSx improved the prediction accuracy in non-European populations, the overall predictions in the non-European populations, especially in the AFR population, remained low relative to the predictions in the EUR population, reflecting the current Eurocentric bias in the discovery GWAS.

Page 8, Line 301-302:
"That said, the absolute prediction accuracy in the AFR population was low relative to the predictions in the EUR and EAS populations, because both discovery samples are genetically distant (Supplementary Fig. 8)."

Page 9, Line 343-345:
"We note, however, that the overall prediction accuracy in the AFR population remained low relative to the predictions in EUR and EAS individuals, reflecting highly imbalanced sample sizes in the training GWAS across populations (Supplementary Fig. 8)."

Page 13, Line 504-510:
"Lastly, we note that although PRS-CSx can improve cross-population polygenic prediction, the gap in the prediction accuracy between European and non-European populations remains considerable, and many predictions in non-European populations are not practically useful. Indeed, sophisticated statistical and computational methods alone will not be able to overcome the current Eurocentric biases in GWAS. Broadening the sample diversity in genomic research to fully characterize the genetic architecture and understand the genetic and non-genetic contributions to human complex traits and diseases across global populations is crucial to further improve the prediction accuracy of PRS in diverse populations."



Supplementary Figure 8: Relative prediction performance for single-discovery and multi-discovery PRS construction methods using discovery GWAS summary statistics **a,** from UKBB and BBJ, across 33 traits, in different UKBB target populations (EUR, EAS and AFR); **b,** from UKBB and BBJ, across 21 traits, in the Taiwan Biobank (TWB); **c,** from UKBB, BBJ and PAGE, across 14 traits, in different UKBB target populations (EUR, EAS and AFR). Each data point shows the relative increase of prediction performance, defined as $R^2/R^2_{\text{PRS-CS (UKBB)-EUR}} - 1$, in which $R^2_{\text{PRS-CS (UKBB)-EUR}}$ is the R^2 of the trait **in the EUR population** using PRS-CS trained on

the UKBB GWAS summary statistics. In UKBB target populations (panels a and c), R^2 were averaged across 100 random splits of the target samples into validation and testing datasets. The crossbar indicates the median of the relative increase of predictive performance across the traits examined. “median N” indicates the median sample size across the respective discovery GWAS. The trait MCHC was not included in the AFR panel because its R^2 from PRS-CS (UKBB) was almost 0, which inflated relative increase of prediction performance for other methods.

2. Throughout the paper, the authors report the relative increase of performance, however, in figure 3, the ratio of performance were reported instead (R2 / R2PRS-CS (UKBB) instead of (R2 - R2PRS-CS (UKBB)) / R2PRS-CS (UKBB)). It might be best to keep the reported metric consistent.

Thanks for the suggestion. We have now updated Figure 3 and its legend to report the relative predictive performance of each method with respect to PRS-CS trained on the UKBB GWAS, i.e., $R^2/R^2_{\text{PRS-CS (UKBB)}}$ -1.

3. On page 6 line 199-200, the authors stated that “However, when predicting into non-EUR populations, Bayesian multi-discovery methods demonstrated a clear advantage over single discovery methods.” Based on the results, it seems like all multi-discovery methods, including PT-meta and PT-mult also out-perform the single discovery methods most of the time.

Yes, both PT-meta and PT-mult outperformed single-discovery methods most of the time. However, the improvement was marginal in some simulation settings and was much smaller compared with Bayesian multi-discovery methods. We have revised the text to make the description more accurate (Page 6, Line 206-208):

“However, when predicting into non-EUR populations, multi-discovery methods clearly outperformed single-discovery methods, with Bayesian methods (LDpred2-mult, PRS-CS-mult and PRS-CSx) demonstrating a larger advantage over PT-based methods.”

4. If I understand correctly, to use PRS-CSx, we will first perform the PRS-CSx analysis to obtain the adjusted summary statistics for k populations across a few \phi. To obtain the PRS-CSx PRS, we then need to fit the k PRS in a linear regression model (or logistic regression for binary trait), and then obtain the individual coefficients. The “final” PRS is then calculated by applying the coefficient from the linear regression to the k PRS calculated for the validation data set. If this is the case, will it be possible for us to use PRS-CSx in a relatively small cohort where split in half analyses might not be viable due to power, or if we would like to obtain PRS for all

samples within our cohort? This might be a common scenario as sample size for the under-represented population is usually small.

This is a great question. The reviewer is correct that, same as PT-mult, LDpred2-mult and PRS-CS-mult, the use of PRS-CSx requires a validation dataset to tune hyper-parameters and learn the optimal linear combination of population-specific PRS, and an independent testing dataset where the final PRS can be generated and evaluated. We expect that, as non-European genomic resources continue to grow, reserving a relatively small number of samples for validation will become increasingly manageable. However, we do recognize that currently some target cohorts may be too small to be split into validation and testing datasets, which limits the use of PRS-CSx and the linear combination strategy in general. To address this issue:

1. We have released the posterior SNP weights (see the github repository for PRS-CSx: https://github.com/getian107/PRScsx) and linear combination weights of PRS-CSx for all the traits and target populations examined in this study (Supplementary Tables 12 and 15) to facilitate the use of PRS-CSx. These weights can be directly applied to the target cohort to generate the final PRS, eliminating the need of a validation dataset.
2. External individual-level datasets that have matched phenotype and ancestry with the target cohort can be used as the validation sample. For example, UKBB included >7,000 African, >2,000 East Asian and >8,000 South Asian individuals, sufficient to be used as an independent validation dataset for a range of complex traits and disease phenotypes. Our analysis in the Taiwan Biobank (TWB) used this approach, where hyper-parameters and linear combination weights were optimized in the UKBB EAS sample, and applied without modifications to the TWB sample.

We further note that, while these efforts and strategies may partially resolve the issue, as non-European genomic resources remain limited, for some phenotypes of interest, independent validation and testing datasets may be difficult to identify, and PRS evaluation may rely on a single small target cohort. In addition, in certain applications, as the reviewer pointed out, it may be preferable to calculate PRS for all samples within the target cohort rather than stratifying them into different ancestry groups. For example, returning genomic predictions to recently admixed patients in clinical settings would be difficult as ancestries are not distinct entities, and genetic ancestry assignments may be inconsistent with self-reported race/ethnicity, illuminating the complexity of communicating population-stratified PRS results to patients. In these scenarios, PRS-CSx provides an “auto” version which automatically learns the global shrinkage parameter from the discovery summary statistics, and a “meta” option which integrates population-specific posterior SNP effects using an inverse-variance-weighted meta-analysis within MCMC iterations. Combining the “auto” and “meta” algorithms thus generates a trans-ancestry PRS that can be applied to all samples in the target cohort without the need for a validation dataset. We note that, although simpler to implement, the “meta” option is expected to be less accurate compared with the linear combination approach that optimizes PRS estimation

separately in each target population. In this revision, we have added new text to discuss the above options that can deal with small target cohorts, as well as their advantages and limitations (Page 12, Line 445-466). As the goal of the manuscript is to maximize the prediction accuracy of PRS in major population groups, we did not include assessment of the “auto” and “meta” algorithms, but saved the application and assessment of this trans-ancestry PRS construction approach to separate work that has a focus on the clinical translation and implementation of PRS (medRxiv preprint: https://doi.org/10.1101/2021.09.11.21263413).

5. Another question I have is regarding non-overlapping SNPs. On page 14 line 538-542, it stated that monomorphic or rare variants do not present in the population specific LD panel for population A are not included in the construction of PRS for population A, but if the variants were missing in other population but not in population A, it will be included in the PRS calculation but not coupled in the posterior inference. Does that mean that for any SNPs that were only presented in one of the populations, the effect size of those SNPs will be applied as is without any shrinkage? To help me better understand this, imagine an extreme scenario where we have two populations A and B and the genotyping chips of these two populations does not share any SNPs. In this unlikely scenario, is it safe to assume that PRS-CSx will reduce into PRS-CS, where each of the populations were essentially analyzed separately and have their own adjustment?

For SNPs that are only present in one population but missing in other populations, a shrinkage is still applied to the marginal effect size of the SNP, but the amount of shrinkage is not coupled across populations. The reviewer is correct that in the extreme scenario where there is no overlapping SNP between input GWAS summary statistics, PRS-CSx reduces to applying PRS-CS separately to each discovery GWAS. We have clarified this in the text (Page 15, Line 597-599):

“In the extreme, unlikely scenario, where there is no overlapping SNP between input GWAS summary statistics, PRS-CSx reduces to applying PRS-CS separately to each discovery GWAS.”

Reviewer #3

Remarks to the Author:

The revised manuscript by Ryan et al present new simulations and analyses that clarify and improve the manuscript. I thank the authors for the detailed comments with my remaining comments being on text clarifications.

1. I could not find the derivation for the statement that under the proposed model the posterior mean effects is (D_k + T^-1)^2\beta_hat_k. In particular, the equation describing \beta_jk from main text is different from the supplementary. It appears \phi could be propagated through the gamma distributions to reach similar equations, but why the different equations? Also, no derivation is provided for the MCMC equations being correct to sample from the posterior of \beta as defined in the model. I encourage the authors to tighten up their presentation of the mathematical aspects of their approach such that the reader can replicate/understand the mathematical details of their model.

We apologize for not providing sufficient mathematical details for the model and posterior inference. The marginal density of beta is equivalent whether the global shrinkage parameter (ϕ) scales the variance of the normal distribution or the rate parameter of the gamma distribution. We presented the former formulation in the main text because we thought it would be easier to see how beta scales with both the local and global shrinkage parameters. To avoid confusion, we now present consistent models in the main text (Page 15, Line 573) and Supplementary Information, where ϕ is included in the gamma distribution (which is the model we implement in posterior inference), but also clarify that the variance of beta scales with both the local and global shrinkage parameters (Supplementary Methods). In addition, we now provide in the Supplementary Information a step-by-step derivation of the full conditional distributions used in the Gibbs sampler (Supplementary Methods). We hope that these additions provide sufficient details of the mathematical aspects of the PRS-CSx model.

2. I commend the authors for the greatly expanded simulations exploring various parameters and impact of architecture on their results (Supp Figs 1-7).

Many thanks for the positive comments!

3. The authors continue to present their results in the main text conflating sample size vs PRS method in assessing improvements in PRS performance. While it is ultimately the authors choice on how to present their main results, I continue to find this style of presentation likely to lead to misleading interpretations.

We thank the reviewer for the continued discussion on this important topic. We fully agree that it is critical to deliver our results with clarity to prevent potential misleading interpretations. With justifications provided in the next paragraph, we have now added the median sample size of the discovery GWAS to the legend of both Figure 2 and Figure 3 such that readers can directly appreciate the fact that single-discovery and multi-discovery methods are trained on discovery GWAS with different sample sizes.

We recognize that given the various factors that can influence prediction accuracy (e.g., different methods, target populations and training sample sizes), and the massive information that can be extracted from the analysis (e.g., comparison of different PRS methods, comparison of prediction accuracy across target populations, and comparison of the predictive power of different training GWAS), each presentation style has their limitations. We fully agree with the reviewer that the current presentation style is not perfect. As we discussed in our previous response letter, while aligning the total training sample size reveals the contributions of factors other than the sample size to the prediction accuracy, our current presentation has the benefit of aligning with the common practice in the field and best summarizing the information we would like to convey in the main text. Specifically, we chose this analysis design because (i) it mimics the real-world scenario where each method is applied to the largest GWAS available; (ii) it shows the advantage of integrating data from multiple populations over single-discovery methods in cross-population prediction.
Upon the reviewer's comment, we have clarified the sample size in the figure legend. We hope this provides additional clarity and reduces the chance of misinterpreting the results.

Decision Letter, second revision:

Our ref: NG-A56583R2

29th Dec 2021

Dear Dr. Huang,

Thank you for submitting your revised manuscript "Improving Polygenic Prediction in Ancestrally Diverse Populations" (NG-A56583R2). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Congratulations on the paper!

All the best,

Catherine

Catherine Potenski, PhD
Chief Editor
Nature Genetics
1 NY Plaza, 47th Fl.
New York, NY 10004
catherine.potenski@us.nature.com
https://orcid.org/0000-0002-4843-7071

Reviewer #2 (Remarks to the Author):

With the latest update of the manuscript, most of my concerns are now addressed. Thank you to the authors for their hard works. I have just one final comment.

In recent analysis, I tried to replicate the authors' simulation procedure to generate population stratified data using HapGen. While the resulting genotype data does have the expected MAF structure of their corresponding population, the actual population structure were not representative e.g. Fst between the European and African samples were much closer than expected (less than 0.1) and the different populations were perfectly separated on PC1. This should not be a serious problem for the authors as they have also conducted many analyses on real data, however, given that there are a relatively large fraction of results were conducted on the simulated data, it might worth putting this problem down as a possible limitations. Other than that, I have no other comments and I think this paper is ready for publish.

Thank you for the hard works.

Author Rebuttal, second revision:

NG-A56583
Improving Polygenic Prediction in Ancestrally Diverse Populations
Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Stanley Global Asia Initiatives, Lin He, Akira Sawa, Alicia R. Martin, Shengying Qin, Hailiang Huang, Tian Ge

Reviewer #2:
Remarks to the Author:
With the latest update of the manuscript, most of my concerns are now addressed. Thank you to the authors for their hard works. I have just one final comment.

In recent analysis, I tried to replicate the authors' simulation procedure to generate population stratified data using HapGen. While the resulting genotype data does have the expected MAF structure of their corresponding population, the actual population structure were not representative e.g. Fst between the European and African samples were much closer than expected (less than 0.1) and the different populations were perfectly separated on PC1. This should not be a serious problem for the authors as they have also conducted many analyses on real data, however, given that there are a relatively large fraction of results were conducted on the simulated data, it might worth putting this problem down as a possible limitations. Other than that, I have no other comments and I think this paper is ready for publish.

Thank you for the hard works.

We thank the reviewer for pointing this out. We have revised our manuscript accordingly by adding: “We note, however, that while highly scalable, genotypes simulated by HAPGEN2 may not fully capture the complex population structure within and across ancestry groups.”

Final Decision Letter:

In reply please quote: NG-A56583R3 Huang

16th Mar 2022

Dear Dr. Huang,

I am delighted to say that your manuscript "Improving Polygenic Prediction in Ancestrally Diverse Populations" has been accepted for publication in an upcoming issue of Nature Genetics.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Genetics style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf,

and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-A56583R3) and the name of the journal, which they will need when they contact our Press Office.

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that Nature Genetics is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that Nature Research offers an immediate open access option only for papers that were first submitted after 1 January, 2021.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in natureprotocols.com, you are enabling researchers to more readily reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to https://protocolexchange.researchsquare.com/. After entering your nature.com username and password you will need to enter your manuscript number (NG-A56583R3). Further information can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#protocols

Congratulations to you and your team on this paper!

All the best,

Catherine

Catherine Potenski, PhD
Chief Editor
Nature Genetics
1 NY Plaza, 47th Fl.
New York, NY 10004
catherine.potenski@us.nature.com
https://orcid.org/0000-0002-4843-7071